CDF Offline Report

Donatella Lucchesi
INFN-Padova, U. of Padova

Rick Snider
Fermilab

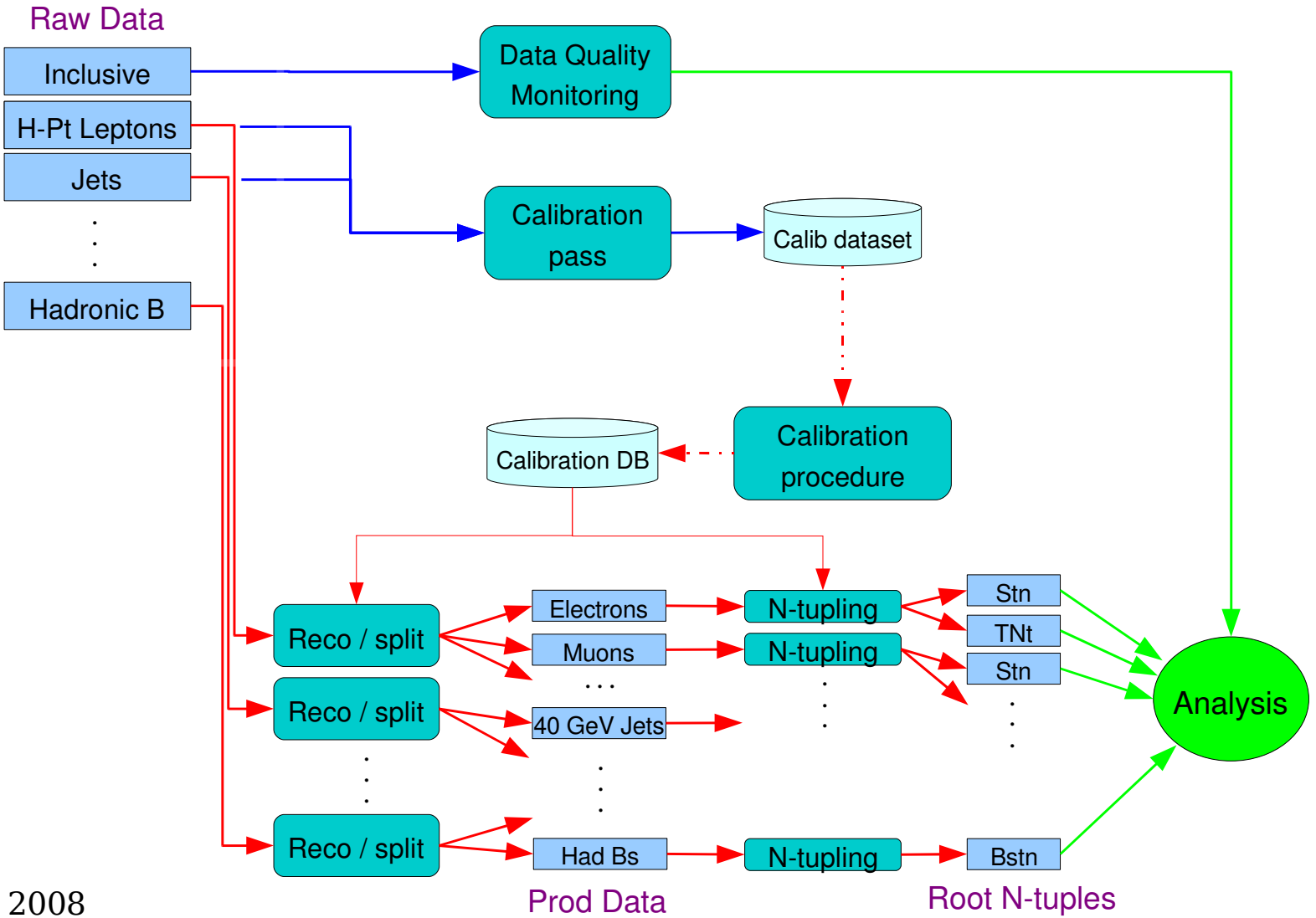
Raw data production model
Production operations
Monte Carlo data production
Analysis Computing

All Experimenter's Meeting
October 27, 2008

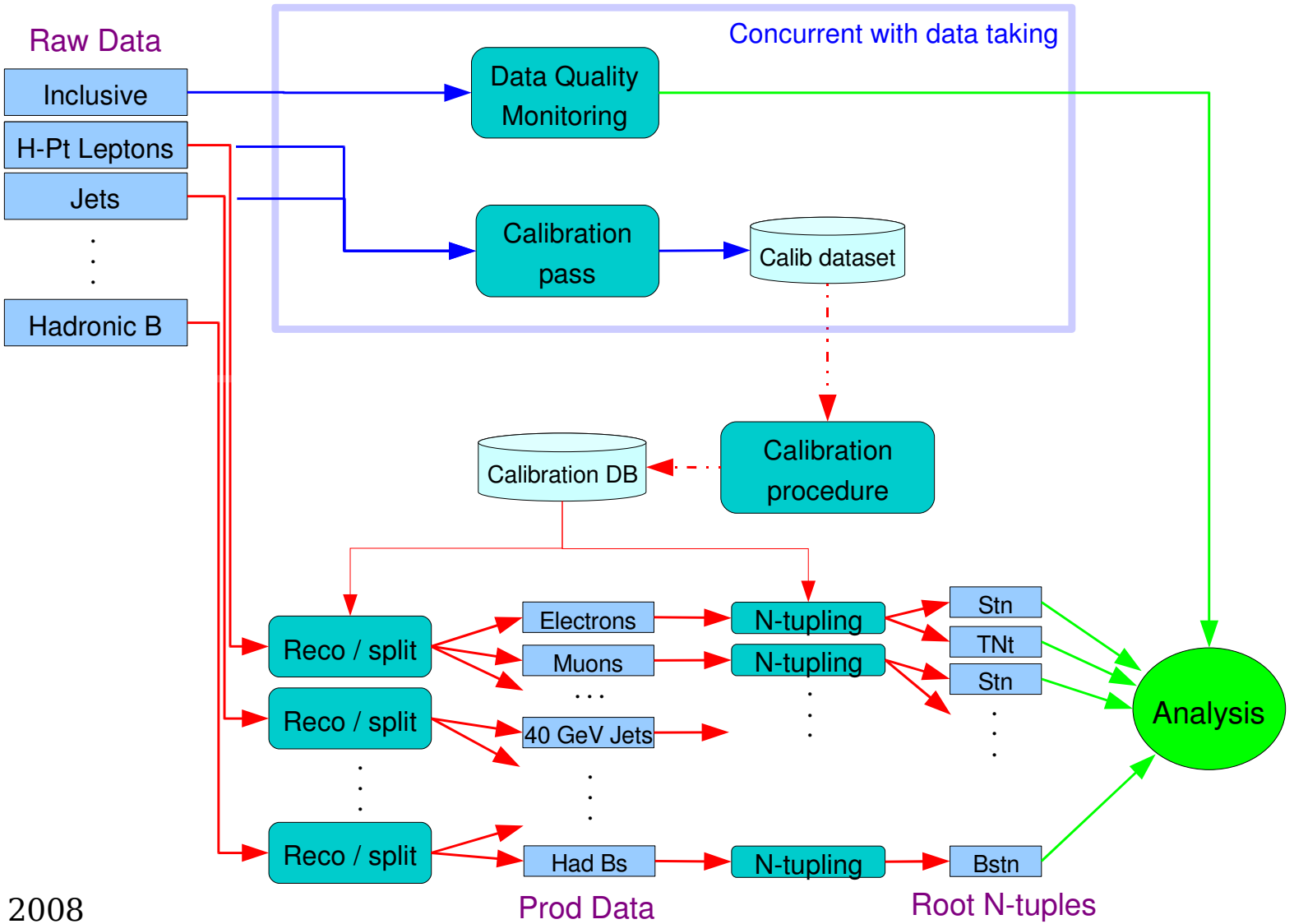
Raw data production model

- Goals of offline production operations
  - Deliver data required for analysis as close to data taking as possible
    - Final compressed datasets from reconstructed raw data
  - Ensure production is not the limitation in the rate of physics output
- The processing problem
  - Log data at rate of 5 – 7 M events/day
  - Calorimeters require re-calibration every ~3 months
    - Need to accumulate ~150+ M events to calibrate (though not all used for calib)
- Strategy
  - Divide data into “run periods” of 4 – 10 weeks
    - Typically 200 – 400 M events
  - Process data by run period
    - Calibration, raw data reconstruction, ntuple creation
  - Analyses use multiple run periods as needed for new results

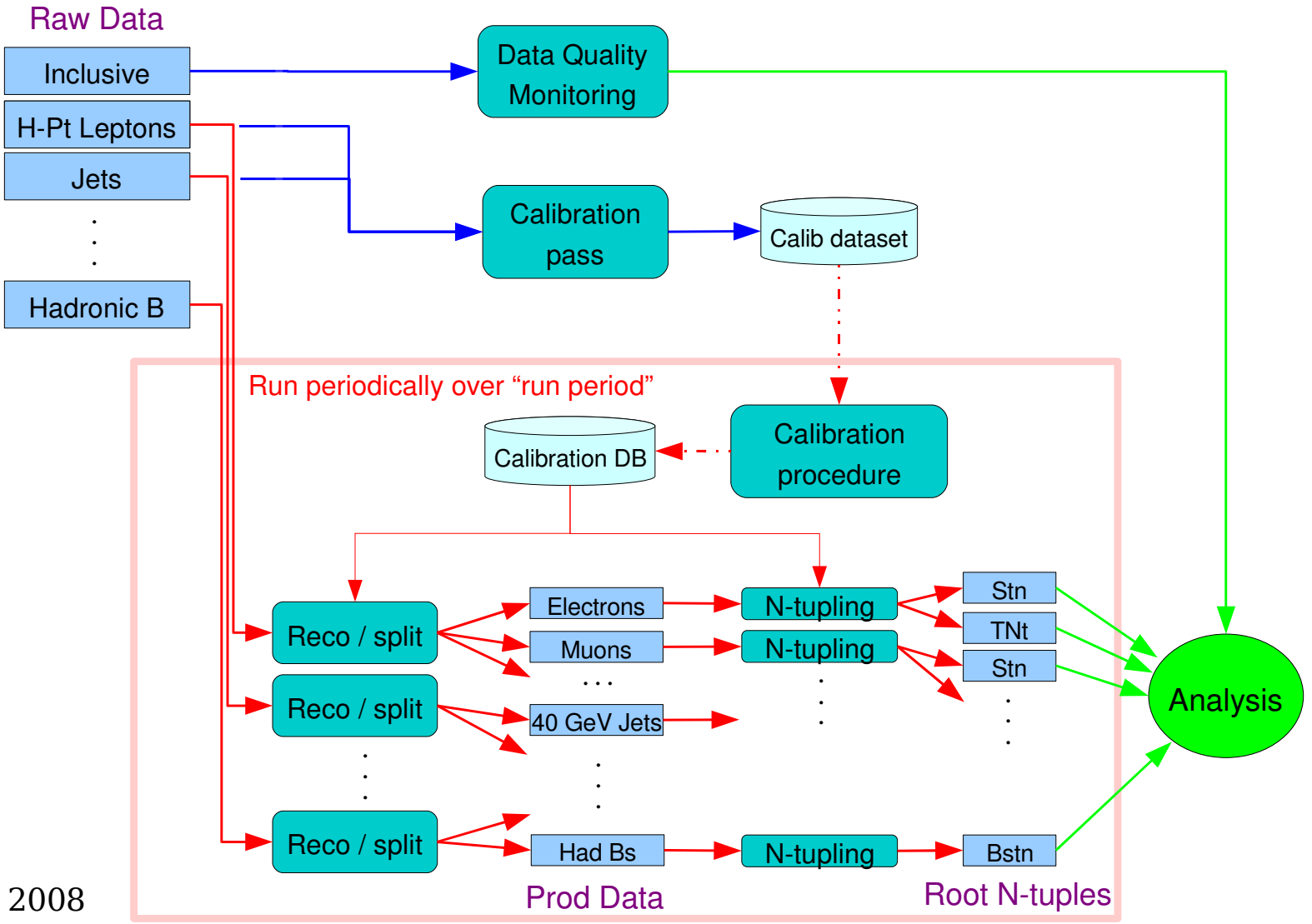
Raw data production model



Raw data production model



Raw data production model



The production cycle

- Detector calibrations
  - Process about 30% of the raw data within a few days of data taking
  - Calculate calibrations and perform validation for each run period

  Typically completed 3 – 6 weeks after end of run period
- Raw data production
  - Reconstruction of data
  - Split data into datasets into physics datasets based upon triggers
    - 42 full + 9 compressed datasets

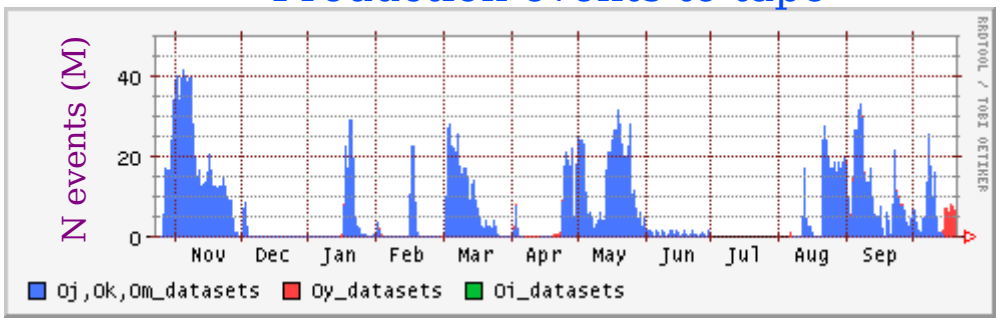
  Typically completed 3 – 6 weeks after calibrations ready
- Ntupling
  - Performed on production output (after splitting)
    - Prioritize processing to do most important first
  - Three partially overlapping flavors: standard, top, Bs

  Typically 2 – 3 days behind raw data production

Raw data operations

- Event reconstruction
  - Average processing time
    - ~2 sec/event across all streams and luminosities (varies greatly event type)

Production events to tape



Data delivery for recent run periods

Period	Start	End	Lum (pb-1)	Events (M)	N-tuples ready
13	May 13, 07	Aug 4, 07	317	545	Nov 29, 07
14	Oct 28, 07	Dec 3, 07	45	59	Feb 21, 08
15	Dec 5, 07	Jan 27, 08	159	210	Apr 7, 08
16	Jan 27, 08	Feb 27, 08	142	168	May 21, 08
17	Feb 28, 08	Apr 16, 08	188	235	Jun 6, 08
18	Apr 18, 08	Jul 1, 08	407	436	Oct 25, 08

10 – 12 weeks for most

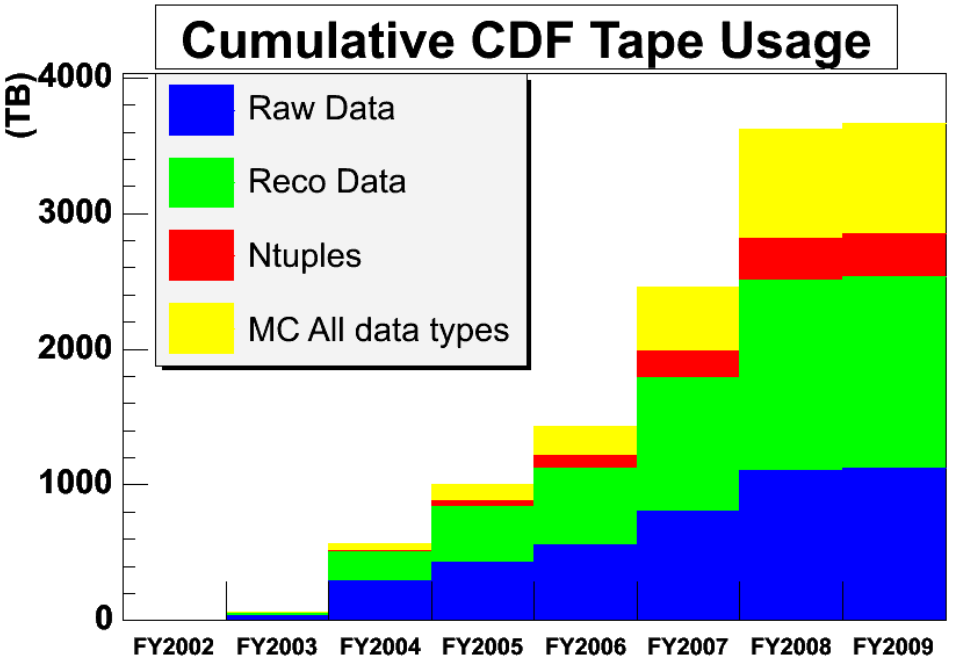
Continue to work on improvements to address rate limitations

Raw data operations

- Data processed on-site
  - Past run periods processed on 600 node farm dedicated to CDF
    - Also used for calibrations, N-tupling and analysis
  - Currently migrating processing to Fermigrid-based farms
    - Final stage of migrating all CDF computing into Fermigrid
      - Better optimizes CPU utilization
    - All processing for the next run period will be performed on Fermigrid
- Data re-processing
  - About 30% of data is processed twice as part of production cycle
    - Once for calibrations, once for physics datasets
  - The experiment has no plans for large scale re-processing

Data volumes

- Data on tape
  - Total of 3.6 PB
  - Raw data
    - 7.9 billion events
  - Monte Carlo data
    - 4.6 billion events
    - Includes a combination of centrally produced MC and analysis-specific MC

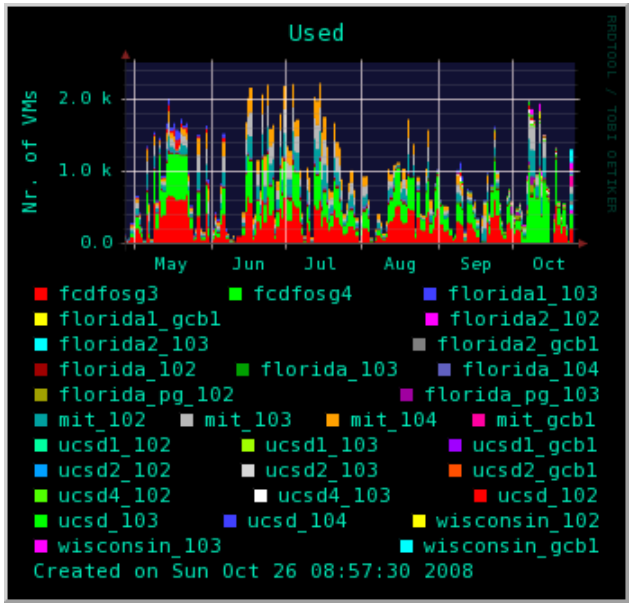


Monte Carlo data production

- The “old” MC data production model
  - Run-based MC that takes into account detector configuration and luminosity
  - Required continuous MC production operations coordinated with data taking
- Changing the production model for new MC
- The new MC production model
  - Luminosity profile scaling
    - Generate MC asychronously with data taking
    - Allows better scheduling of CPU usage
    - Significantly reduces amount of MC needed relative to run-based approach
  - Possible because the detector configuration is very stable

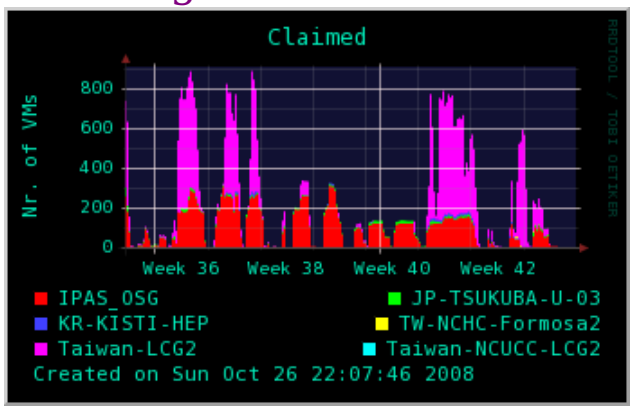
Monte Carlo data production

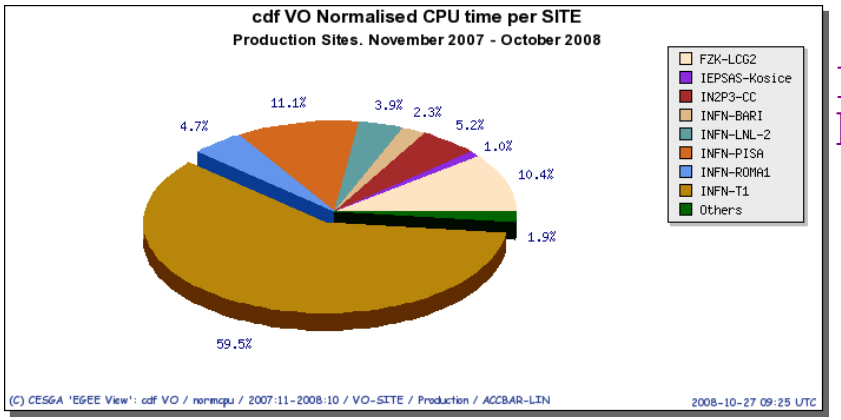
- Centralized MC produced off-site
  - Open Science Grid
    - US institutions
    - Same technology for Pacific Rim
  - LHC Computing Grid
  - INFN-CNAF
    - Priority access to CNAF T1
  - Barcelona



OSG usage by site + farm

Pacific Rim usage by site + farm

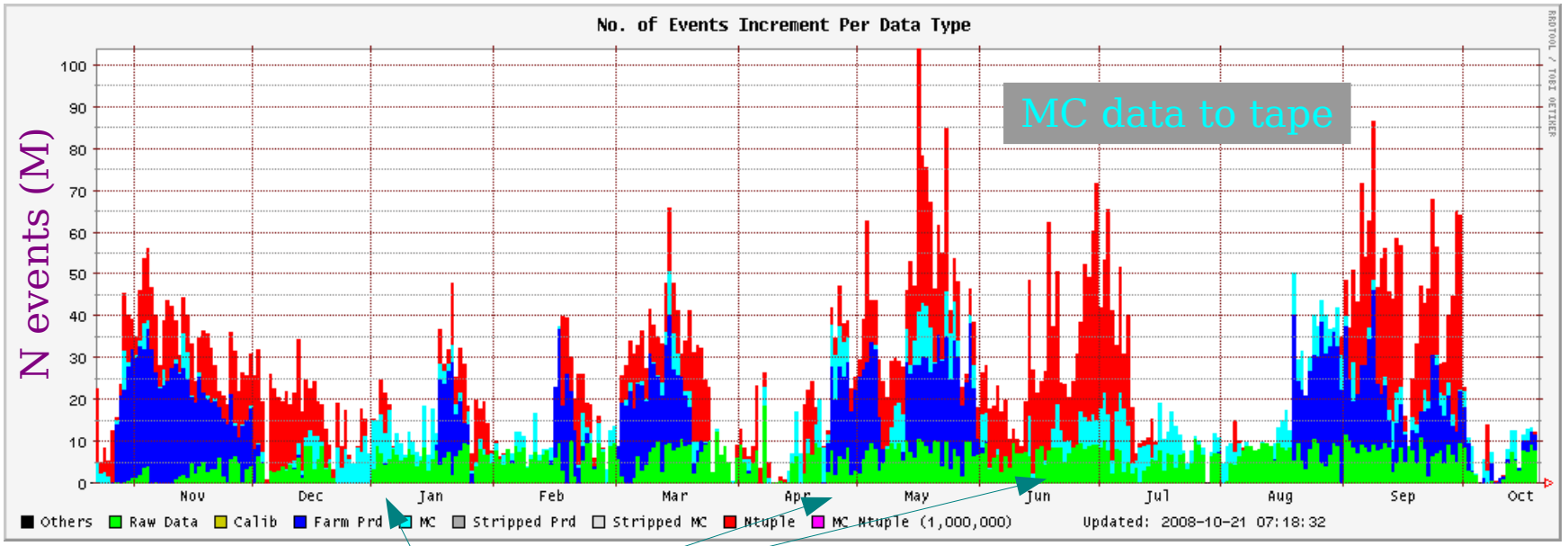




LCG usage by site

MC data production operations

- MC data generated
  - 1.1 G events produced last year
  - Some periods of concentrated production during “MC attacks”



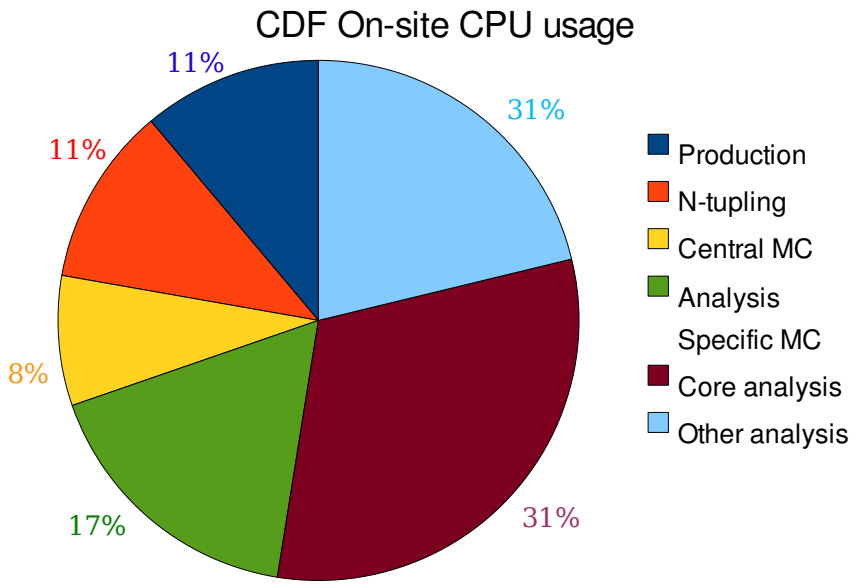
MC attacks

Analysis computing

- Computing requirements scale with:
  - Full data set size
  - Complexity of analyses
  - Number of people / analyses

Computing problem becomes harder with time

- Facilities
  - 5k CPUs on-site for data intensive analysis
    - Shared with production activities
    - Some large datasets also located at INFN-CNAF
  - Off-site computing also available for CPU intensive analysis
    - Matrix element analysis, pseudo-experiments, etc.



Analysis computing

- Is it all effective?
- The bottom line is the physics that CDF produces
  - 50+ new results at 2008 Winter conferences
  - Another 50+ new results at 2008 Summer conferences
  - Expect ~40 publications in 2008

Summary

- The CDF offline is successfully meeting the physics needs of the experiment
  - Due to the hard work of many collaborators at Fermilab and around the world
  - A close and productive collaboration with the Computing Division has been critical to this success

    "Thank you" to the CD!
- Will ensure continued success by working to improve the systems, increase efficiency and reduce the effort required to conduct computing operations.